Final Project Report

DOE ECPI Award: DE-FG02-04ER25610
Title: Transfer Function Design for Scientific Discovery

Length of Project: 10/2004-05/2008
(including a 9-month no cost extension)

Principal Investigator: Jian Huang
Contact: 203 Claxton Complex
University of Tennessee, Knoxville, TN 37996
Phone: 865-974-4398
Email: huangj@cs.utk.edu

1. Participants

During the project, PhD students: Markus Glatter, C. Ryan Johnson, Joshua New and Wesley Kendall, have been supported as Graduate Research Assistants (GRA). Among them Markus Glatter will graduate with a PhD in December 2008. The PI also supported a MS student, Colin Mollenhour, as a GRA on a short-term basis during Year 2. All students are in the computer science graduate program at the University of Tennessee.

The PI collaborates closely with researchers at Oak Ridge National Lab (ORNL) and Spallation Neutron Source (SNS) in the following contexts (in order of effort and foci):

(i) Sean Ahern (Visualization Taskforce Lead at ORNL) and his staff, on collaborative research and on deploying our research software to VisIt and ORNL's visualization production systems.
(ii) Forrest Hoffman, Dr. David Erickson III and Dr. John Drake at ORNL on innovative domain-integrated visualization to support their leading edge climate modeling research.
(iii) Dr. Elissa Chesler, on applying general principles of transfer function design to problems in the broad area of systems biology.
(iv) Dr. Stephen Miller and his staff on integrating our visualization software with the initial production visualization infrastructure at SNS for the neutron science community.
(v) Dr. Tony Mezzacappa and Dr. John Blondin on applying our novel tools to basic astrophysics research. Dr. Gene Ice and Dr. John Budai on designing transfer functions to visualize polycrystalography physics data.

All of the above collaborations have started taking shape since the PI joined the University of Tennessee as an assistant professor of Computer Science. This DOE ECPI grant has provided crucial personnel and resource support for substantial progresses to result from these collaborations.

2. Activities and Findings

Transfer function plays a direct role in determining the data analytic capability of any visualization systems that target volumetric simulations. About five years ago, the field of visualization focused on transfer function design for volumetric medical imaging data and succeeded in discovering a set of effective methods to design transfer functions, most of which are primarily based on various orders of spatial gradients of a single variable. Many of those methods, however, cannot be directly borrowed to study typical simulation data produced by computational scientists. This is because computational data are fundamentally different from medical data and also have different user needs.

Unlike medical data, simulation data do not have boundaries that are as well defined as universal anatomical structures. In addition, while medical imaging such as CT and MRI is often concerned with variables that are independent of each other, leading scientific simulations typically involve multiple interacting time-varying variables, in some cases several hundred of such variables. When visualizing such simulation data, it is ad hoc (in some cases, very tricky) to decide which variable's gradient to use in traditional transfer functions. This element makes it particularly difficult to use medical transfer functions for scientific simulations. Furthermore, medical datasets are often temporally static. In contrast,

computational simulations are highly dynamic in time. Existing approaches of transfer function designs seldom fully consider time as a first class dimension.

Besides simulation data's unique properties, computational scientists have a more exploratory mode of using visualization than in other domains. Hypotheses are quite commonly known but hard to definitively specify. For instance, in climate modeling we are certain that July should be hotter than January in the northern hemisphere. But we are less certain about the exact differences in temperature between those two months. Many scientific research starts from such qualitative understandings.

With crucial funding provided by DOE Early Career PI Award, we undertook a systematic study of transfer function design for the purpose of enabling discovery in large-scale scientific simulations. There are three components in our study: (i) to develop novel quantitative visualization mechanisms to introduce new capabilities into transfer functions, (ii) to develop inherently parallel infrastructure for data reduction and selection in visualization to ensure our expanded transfer function designs remain functional even for large datasets at tera- to peta-scale, and (iii) to develop innovative ways to distill knowledge by creating summary visualization of large problem spaces of partially specified user hypotheses.

For a more comprehensive variety of transfer function, we developed a broadened set of quantitative visualization operators that revolves around correlation and graph properties [3], statistical distribution [5] and regular expressions [4]. These novel operators are also usable in combination with traditional numerical based metrics such as derivatives and integrals. With such a large number of quantitative operators to use, a common theme of our visualization research is to design a set of mini programming language type of interfaces.

Correlation is a common tool for discovering multivariate relationships hidden in data. However with time-varying data involving tens of variables, the correlation space becomes exponential, especially when temporal lags are considered in computing pair-wise variable correlations. To tackle this problem, we devised a set of tools, called *seeGraph*, to quickly discover strong relationships within the exponential correlation space [3]. Statistical distribution is another high-order property of great interests to scientists. The previous difficulties with using distribution for investigative visualization mainly stem from the difficulty to interact with such mathematical functions. Our research led to a very compact and yet powerful query language called *seeDQ* for interacting with large datasets containing per voxel distribution functions [5]. For a user to specify partial hypotheses for visualization, we developed a succinct visualization language interface similar to how regular expressions (regex) are used in textual searches. This mini-language interface is called *seeReg*. The power of *seeReg* roots from its capability to automatically expand on partially specified features and reveal all possible matches in the dataset in a succinct view [4].

Figure 1 shows a visualization of DOE C-LAMP climate modeling data using seeGraph, seeDQ and seeReg together. This visualization is novel in that it allows fuzzy user knowledge to be directly used in creating visualization. This perspective of uncertainty visualization is unprecedented in the field.

While it is already challenging to develop a function-rich visualization methodology, it is an even bigger challenge to build a working system that provides such novel visualization functions on terabytes or petabytes of data. The scalability challenge is un-precedented. Elegant designs of parallelism are indispensable to achieve the necessary scalability.

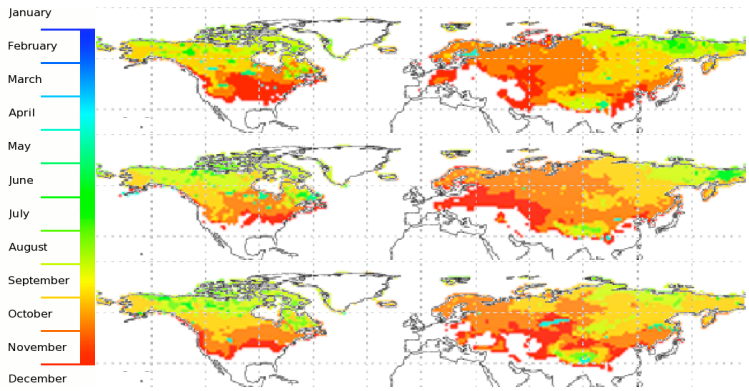

Figure 1. An uncertainty-tolerant visualization created from statistical distribution information using regular expression type of visualization interface. This visualization shows variation in time of first "major" snow fall in the northern hemisphere in a C-LAMP simulation. The years shown from top to bottom are 2050, 2051 and 2052.

The common parallel infrastructure that supports all of our quantitative visualization operators and runtime needs of data reduction is derived from an observation. That is, compound query, albeit straightforward, is indeed the common element fundamental for all runtime visualization data management. For instance, gradient based transfer function designs for medical visualization can all be factored into a set of compound queries. We developed a query-driven parallel visualization system in [1]. The system is independent of grid type, inherently parallel and scalable, and distilled to be state-free in the sense that it operates like web servers (allowing as many concurrent clients as desired [1]). The system not only operates with near-optimal load balance on centralized computing clusters, but also efficiently works in geographically distributed settings [6]. Figure 2 shows our parallel visualization infrastructure supporting 27 concurrent interactive visualization clients on a Powerwall system to study the well-known Terascale Supernova Initiative (TSI) dataset.

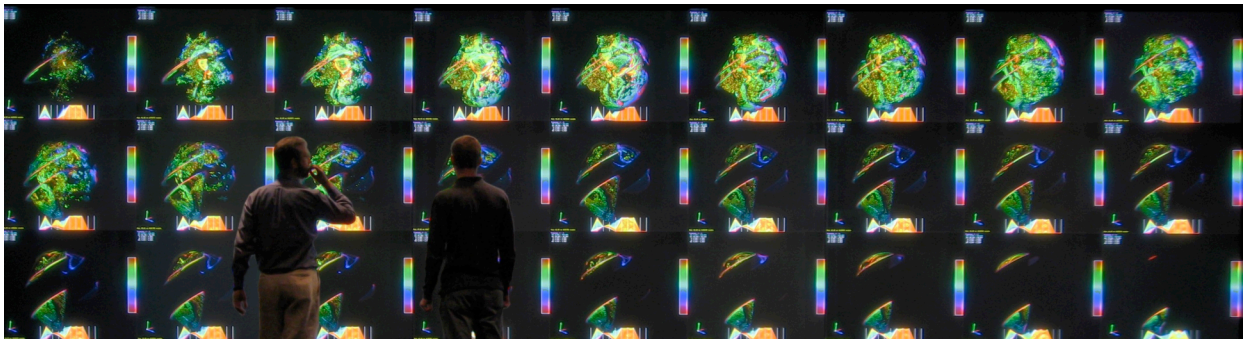

Figure 2. A compound query based concurrent visualization system created to demonstrate the capability of our highly scalable *mqr* system for runtime data reduction and selection. The data shown is the TSI dataset produced by Drs. Blondin and Mezzacappa under the auspices of DOE SciDAC Terascale Supernova Initiative. The visualization shows 27 concurrent interactive visualizations of the TSI datasets on the Everest powerwall system at ORNL. All 27 visualization clients use the same set of 30 backend *mqr* servers. The researchers standing in front of Everest are Sean Ahern (visualization task leader, ORNL) and Colin Mollenhour (graduate student of Dr. Huang).

Our final research component further augments investigative power of scientists by providing summarizing visualizations. This is necessary due to the continuingly increasing complexity of simulation data. It is also necessary due to our broadened set of methods to analyze data. These two factors together result in a greatly increased variety of perspectives to visualize. With a great variety of views beyond what was previously possible, to quickly distill useful knowledge from the many views on hand is highly desired.

To this end, we explored methods to create concurrent visualization of time-varying multivariate data in summarizing views. This research effort is very closely integrated with ongoing research in domain sciences (particularly SciDAC C-LAMP climate modeling researchers Hoffman and Erickson at ORNL). Our early results are already published in a leading global visualization conference [2] and DOE SciDAC Conference [7]. Through collaboration we have found significant ways in which competing climate models differ from one and another. These differences that we have revealed were previously unknown because they are of higher order than previous tools can handle and require feature specification that must tolerate uncertainty. We are currently in close collaboration with our climate scientists to prepare research results for submission to leading scientific journals.

3. Publications and Products

Supported by the DOE ECPI award, we have **five** papers [1-5] (**four** already published and **one** under revision) in top visualization journals and **two** conference papers [6-7].

The papers are listed as the following.

[1] Markus Glatter, Colin Mollenhour, Jian Huang and Jinzhu Gao, “Scalable Data Servers for Large Multivariate Volume Visualization”, **IEEE Transactions on Visualization and Computer Graphics**, Vol. 12, No. 5, pp. 1291-1299, 2006.

[2] Robert Sisneros, C. Ryan Johnson and Jian Huang, ”Concurrent Viewing of Multiple Attribute-

Specific Subspaces", **Computer Graphics Forum (special issue for EuroVis'08 conference)**, vol. 27, no. 3, 2008.

[3] Joshua New, Wesley Kendall, Jian Huang, Elissa Chesler, "Dynamic Visualization of Gene Coexpression in Systems Genetics Data", **IEEE Transactions on Visualization and Computer Graphics**, Vol. 14, No. 5, 2008.

[4] Markus Glatter, Jian Huang, Sean Ahern, Jamison Daniel and Aidong Lu, "Visualizing Temporal Patterns in Large Multivariate Data using Textual Pattern Matching", **IEEE Transactions on Visualization and Computer Graphics**, Vol. 14, No. 6, 2008.

[5] C. Ryan Johnson and Jian Huang, "Frequency Distribution Driven Investigative Visualization of Volume Data", under review (1st round revision), **IEEE Transactions on Visualization and Computer Graphics**.

[6] Wesley Kendall, Markus Glatter, Jian Huang, Forrest Hoffman, David E. Bernholdt, "Web Enabled Collaborative Climate Visualization in the Earth System Grid", **Proc. of International Symposium on Collaborative Technologies and Systems (CTS 2008)**, Irvine, CA, May 2008.

[7] Robert Sisneros, Markus Glatter, Brandon Langley, Jian Huang, Forrest Hoffman, David Erickson III, "Time-Varying Multivariate Visualization for Understanding Terrestrial Biogeochemistry", **Journal of Physics: Conference Series (SciDAC08)**, July 2008.

In addition to research papers, we have two other research software products:

(i) A parallel interactive visualization system *mqr*, capable of handling terabyte level time-varying multivariate datasets using only around 20 computing nodes. It is based on our research paper [1] and has undergone several versions of upgrade. A recent version of our *mqr* system has been integrated with the leading visualization package, VisIt, and is now released as an add-on component of VisIt.

(ii) We have completed a quantitative operator library and compiled it as a component of the Visualization Cookbook Library (vcblib). The entire vcblib is in-house developed and open-source under LGPL license. Currently the vcblib is under beta testing by researchers at ORNL, SNS, North Carolina State University, University of Memphis and Louisiana State University. We have made the first public release online in October 2007. The library can be downloaded from the webpage of our research group at: http://www.cs.utk.edu/~seelab. Documentation of all library functions is also available at the same URL. The documentation is maintained in a mySQL database accessible via PHP. This allows all web access to obtain the latest documentation.